TECHNICAL REPORT

ISO/IEC TR 29127

First edition
2011-07-01

Information technology — System Process and Architecture for Multilingual Semantic Reverse Query Expansion

Technologies de l'information — Processus système et architecture pour l'extension multilinguale des requêtes sémantiques inverses

Reference number
ISO/IEC TR 29127:2011(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, when the joint technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide to publish a Technical Report. A Technical Report is entirely informative in nature and shall be subject to review every five years in the same manner as an International Standard.

ISO/IEC TR 29127 was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 36, *Information technology for learning, education and training*.

Introduction

Learning, Education and Training (LET) in the context of multilingual cultures on a local and global scale can be problematic, especially when learners are proficient in only one language. One of the multilingual problems in a LET environment is how to query LET materials when the requestor cannot understand or is not proficient in the language of the material available.

For example, how does a person who is proficient in French search for, find, and readily understand digital LET materials in Arabic, if the person is not proficient in Arabic? One solution can be found in a process called the Semantic Reverse Query Expander (SRQE). Based on components such as language ontologies, the SRQE process utilizes Java 2 Platform, Enterprise Edition (J2EE) (J2EE)[1)] services that can take a term in one language (source language), expand the term conceptually, translate the expanded terms (into a target language), and perform a query on a targeted foreign language document set. Returns are translated into the language of the requestor. This Technical Report identifies an existing process and architecture used to query foreign language text files.

Technologies and ontologies (i.e. thesauri) for undertaking this kind of matching and expansion operation have been available for some time (e.g. the work of CYC Corp, Global WordNet, Global WordGrid). Valuable lessons have been learned about what such technologies can and cannot accomplish. This Technical Report does not discuss these pre-existing technologies, or describe the improvement or change that the proposed process presented might represent. A particular approach (theory and practice) with respect to the context of difficulties experienced in regard to multilingual equivalencies and translation are presented in Annex C of this Technical Report.

In Clause 3 of this Technical Report, an implementation of the SRQE process is described in a web environment to help clarify the architecture described in Clause 4. Annex A contains possible linkages to ISO/IEC JTC 1, SC 36 projects and future areas of study.

The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that the process described in this Technical Report may involve the use of patents. ISO and IEC take no position concerning the evidence, validity and scope of these patent rights.

The holders of these patent rights have assured ISO and the IEC that they are willing to grant a free of charge license to an unrestricted number of applicants on a worldwide, non-discriminatory basis and under other reasonable terms and conditions to make, use, and sell implementations of the process contained in this Technical Report. In this respect, the statements of the holders of these patent rights are registered with ISO and IEC. Information may be obtained from the companies listed below.

Raytheon Company
Phillip Berestecki
Intellectual Property and Licensing
870 Winter Street
Waltham, Massachusetts 02451-1449
USA

NOTE 1 This Technical Report refers to one particular process or approach for performing reverse semantic queries; there are other approaches and processes that could be developed for these same purposes.

NOTE 2 The process is not dependent on particular database software, protocols, or data sets. Specific components used in the process are an implementation decision.

1) A widely used platform for server programming in the Java programming language.

Information technology — System Process and Architecture for Multilingual Semantic Reverse Query Expansion

1 Scope

This Technical Report identifies an example of a system-based process to index, query, translate, and manage components used in querying and translating documents in multiple foreign languages, enabling learners in learning, education, and training areas to effectively find and share documents on a global scale.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

NOTE For this Technical Report, the following terms and definitions are not considered to be normative. They are informative, and apply only within the context of this Technical Report.

2.1
coordinate term
words that have the same hypernym

EXAMPLE Boat, yacht, and shrimper, all have the same hypernym, ship.

NOTE Adapted from ISO 1087-1:2000, definition 3.2.19.

2.2
entity extraction
process that seeks to locate, classify, and tag atomic elements in text into predefined categories

EXAMPLE Names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

2.3
hypernym
superordinate concept
word that is more generic or broad than another given word

NOTE 1 Another term for a hypernym is a superordinate concept.

NOTE 2 Adapted from ISO 1087-1:2000, definition 3.2.13.

2.4
hyponym
subordinate concept
word that is more specific than another given term

NOTE 1 Another term for hyponym is a subordinate concept.

NOTE 2 Adapted from ISO 1087-1:2000, definition 3.2.14.

2.5
Java servlet
Java programming language objects that dynamically process requests and construct responses

2.6
meronym
constituent part of, or a member of, something

EXAMPLE "Winchester Cathedral" is a meronym of "Church of England".

2.7
nominalization
use of a verb or an adjective as a noun with or without morphological transformation, so that the word can now act as the head of a noun phrase

2.8
word sense
⟨linguistics⟩ one of the meanings of a word

NOTE A dictionary may have over 50 different meanings of the word "play", with each of these having a different meaning based on the context of the word usage in a sentence.

EXAMPLE We went to see the **play** *Romeo and Juliet* at the theater.

The children went out to **play** in the park.

3 Example SRQE Implementation

This clause provides an implemented example of the SRQE process. The simplistic example of the SRQE implementation illustrates a sequence of actions between a learner and the system. The example provided is based on a possible learning assignment made to a learner in producing a report on trucks using foreign language resources. A learner wanting international information in producing a report on trucks could use the system to gather international information related to trucks for possible inclusion in the report. The system can perform a cross lingual query on documents in a number of languages, and translate the documents into the learner's native language. The learner can optionally access a map of the locations listed in the text files returned by the query for improved comprehension of where the article originates from, or the location of where the subject of the article can be found.

In this example, the SRQE process utilizes a user interface, interacting with Java servlets to provide a web based User Interface (UI). The SRQE process flow is illustrated below in Figure 1. The SRQE process is a human-machine interactive process to perform cross-lingual queries. Human inputs are shown in the white box above the green arrow. Java servlet functions are shown below the green arrow. A UML diagram showing the Java servlets in a green box are linked to functions in the white box.

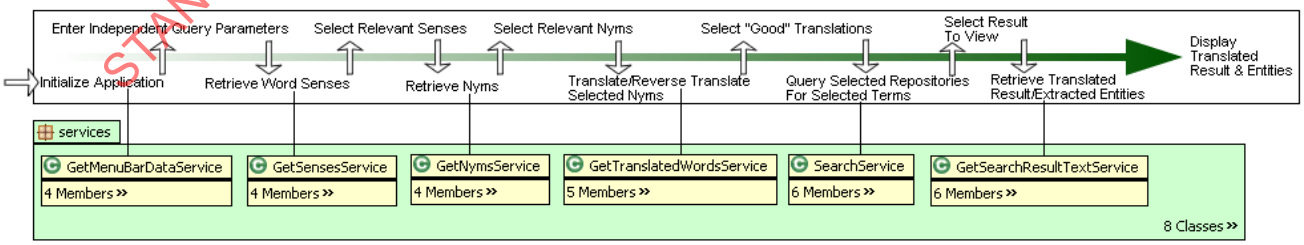

Figure 1 — SRQE Process Flow

3.1 Initialization of the User Interface

The user interface application is accessed by URL. The GetMenuBarDataService initializes the user interface application providing the repositories available and in what languages the repositories are in. The user interface application is shown in Figure 2. The GetMenuBarDataService builds the menus and menu choices in the user interface. This includes 1. Enter word to be translated, 2. Select Class, 3. Select Language(s), and 4. Select Sources, 5. Execute Query, the results section (blank), and the Map.

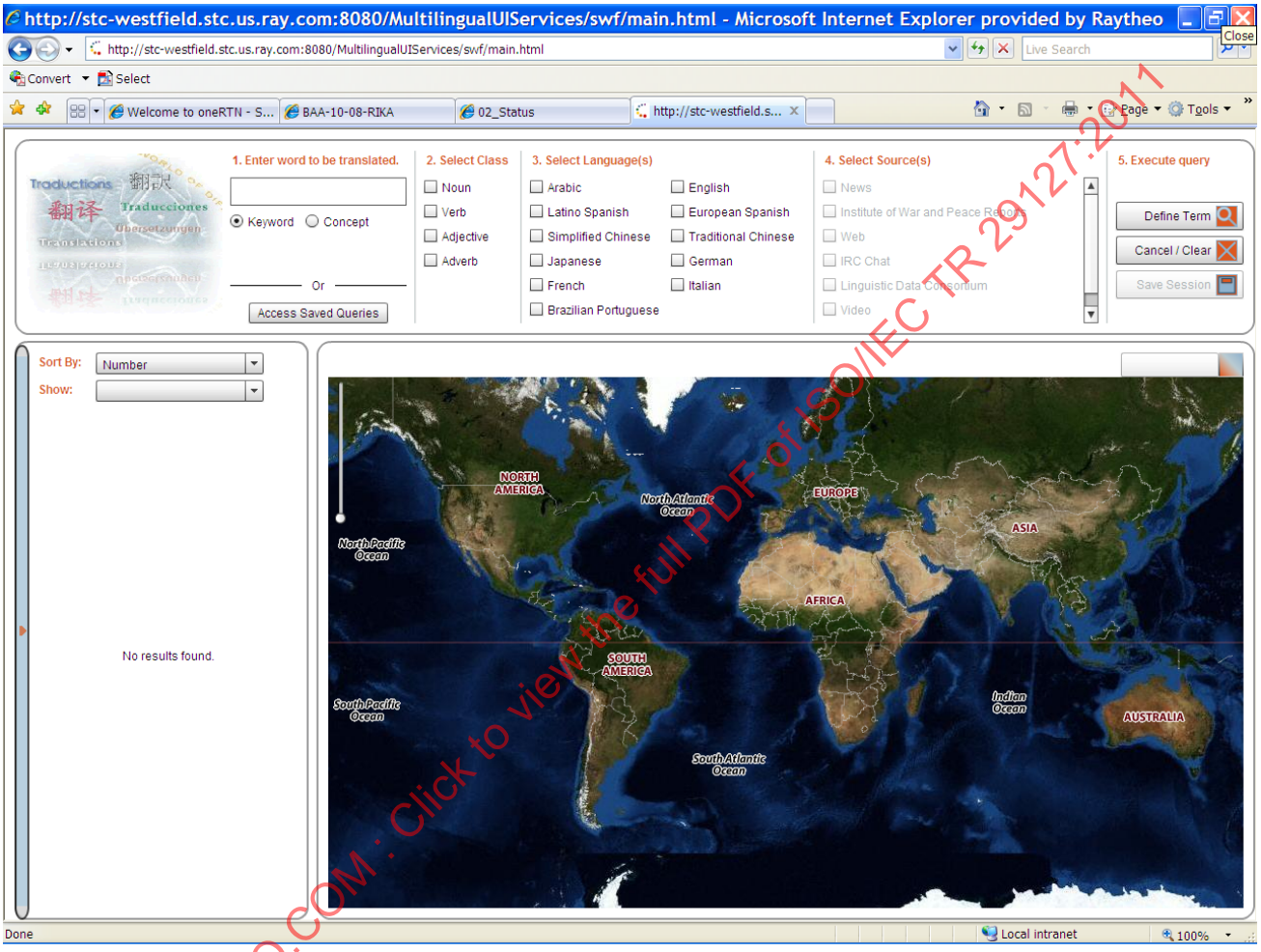

Figure 2 — Initialized User Interface

NOTE 1 In this example, an Adobe Flash plug-in is required for the use of the SRQE in a browser. The user interface described in the example is based on Adobe Flex.

NOTE 2 The original example uses CaMel CaSe format in this instance and the instances that follow. This formatting is retained for this reason.

3.2 Select Query Parameters

The learner enters a word in “1. Enter word to be translated”. The learner selects a class of the word in “2. Select Class”. The learner selects a language in “3. Select Language(s)”. The learner selects a source in “4. Select Source(s)”. In this example, the learner enters Truck in “1. Enter word to be translated”. The learner clicks on Noun in “2. Select Class”. The learner selects Arabic in “3. Select Language(s)”. The learner selects Linguistic Data Consortium as the source in “4. Select Sources”. Figure 3 shows menu items 1 through 4 filled in by the learner.

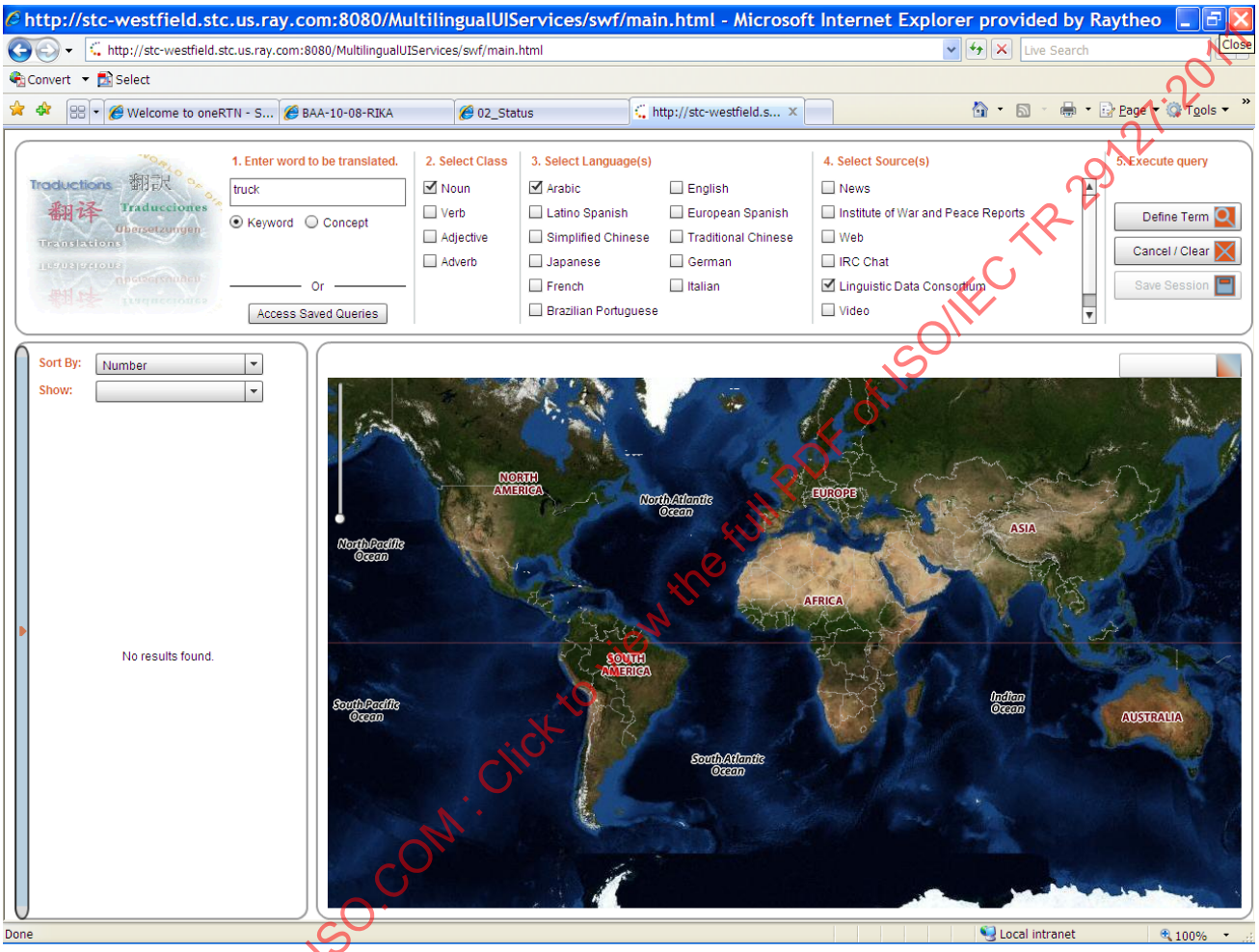

Figure 3 — Learner Inputs Menu Items 1-4

The learner selects Define Term in Item 5 “Execute query” menu. The GetSensesService retrieves the noun word senses for Truck, in this example, word senses are retrieved by GetSensesService from WordNet utilizing the Java WordNet Library (JWNL). The GetSensesService returns noun word senses for truck. The word senses for Truck is displayed in the return section of the interface. Figure 4 shows the noun word senses for Truck returned by the GetSensesService.

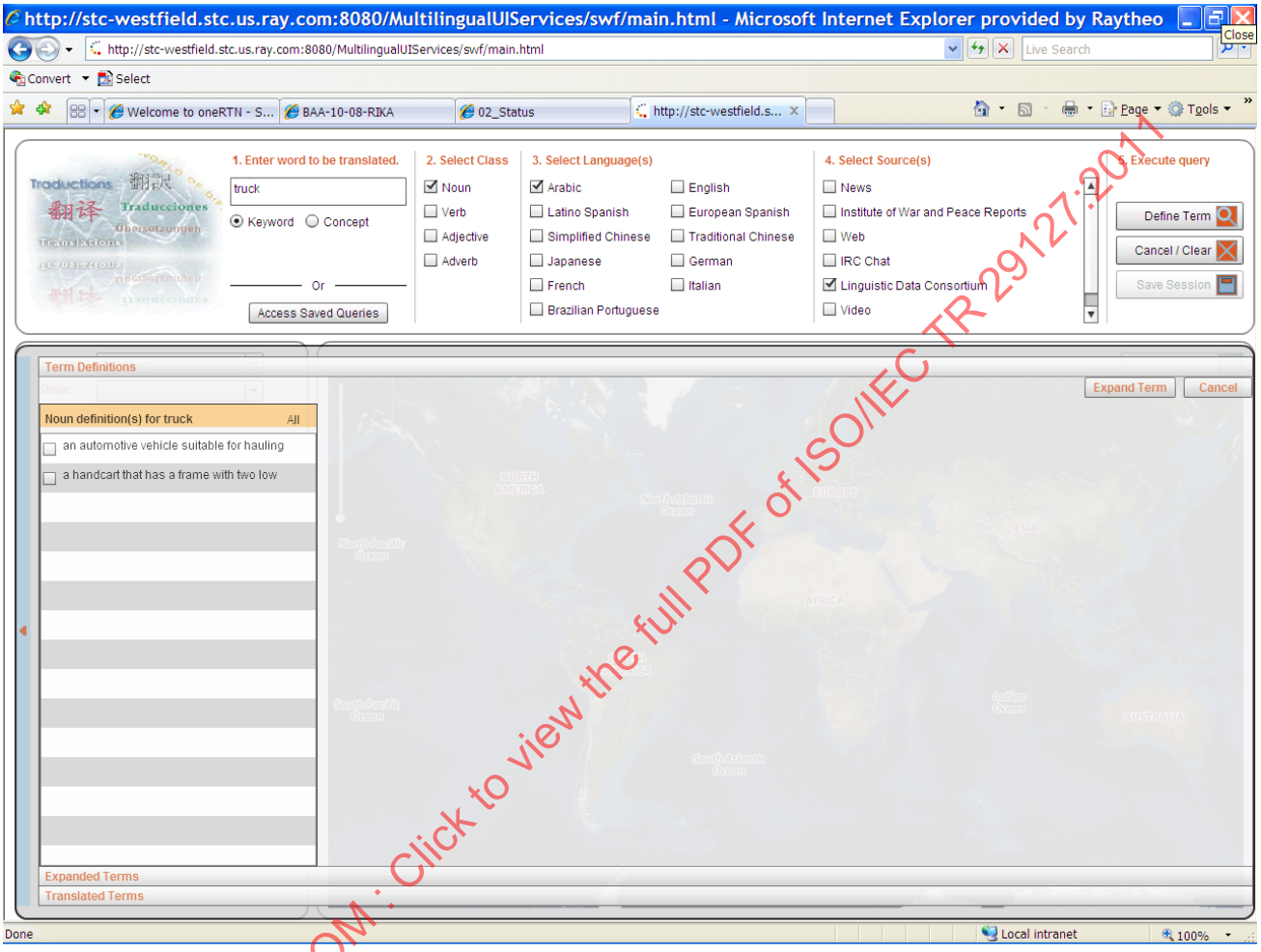

Figure 4 — Word Senses for Truck

3.3 Select Word Senses

The learner selects the relevant word sense for Truck. Figure 5 shows the word sense selected by the learner "an automotive vehicle suitable for hauling".

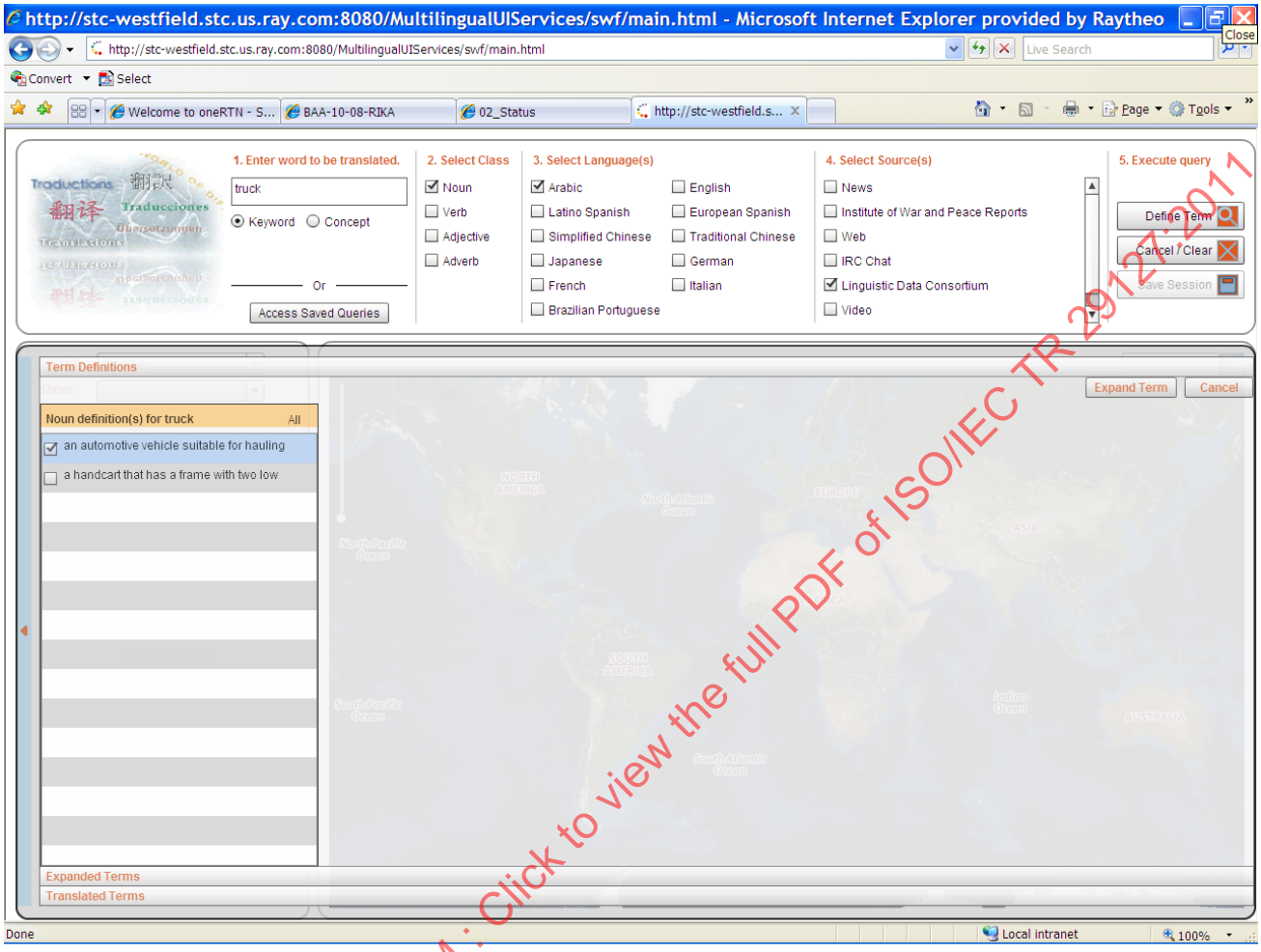

Figure 5 — Word Sense Selected

The learner selects “Expand Term” at the far right of the word sense display. The GetNymService retrieves coordinate terms, hyponyms, nominalizations, hypernyms, and meronyms for the word sense selected. In this example, the GetNymService retrieves expanded terms from WordNet utilizing the JWNL. Figure 6 shows the expanded terms returned from the GetNymService.

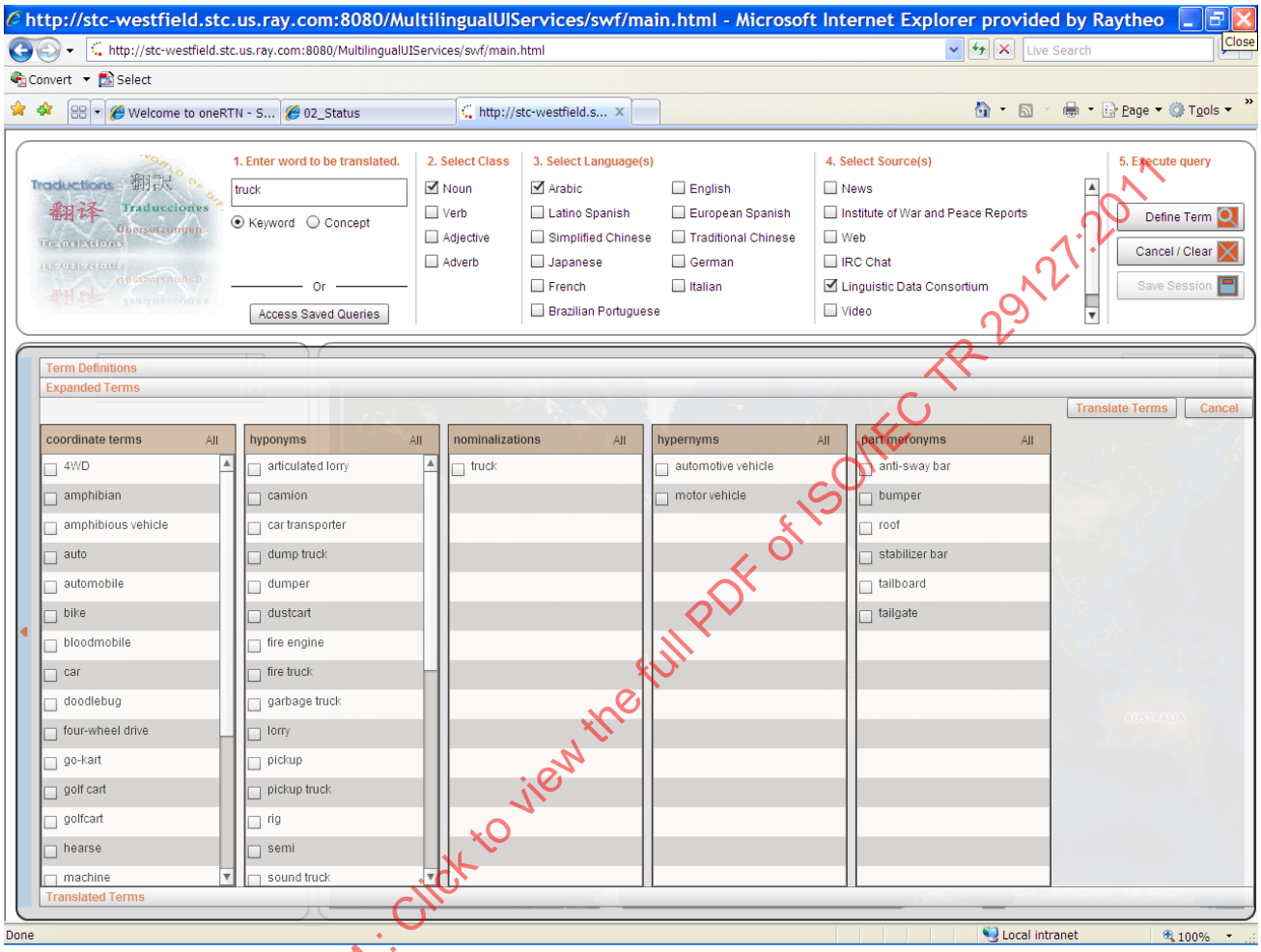

Figure 6 — Expanded Terms List

3.4 Selecting and Translating Appropriate Terms

The learner selects the terms of interest for translation and reverse translation. Terms selected are shown in Figure 7.

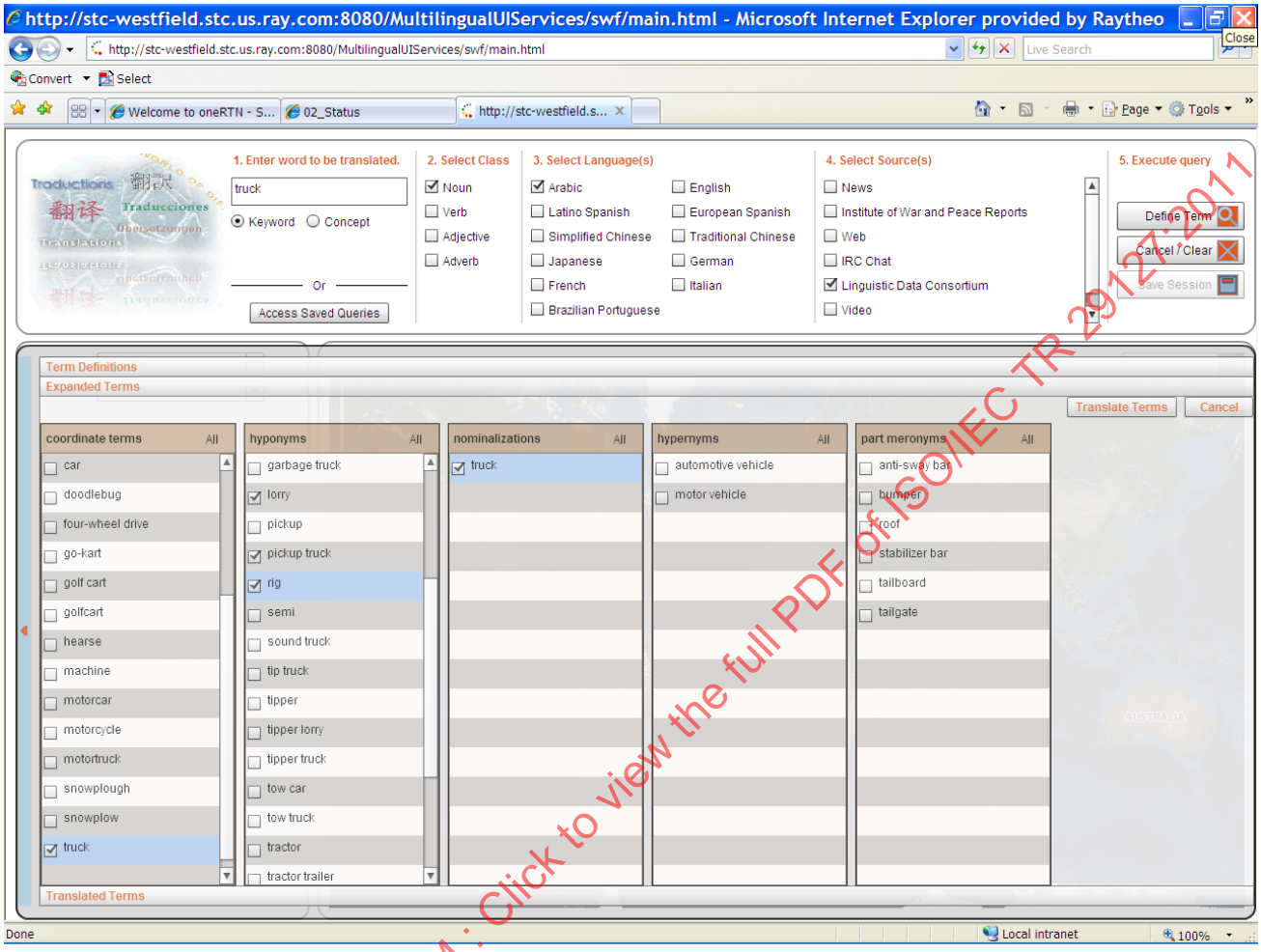

Figure 7 — Terms Selected

The learner selects “Translate Terms” above and to the right of the part meronyms list. The GetTranslatedWordService sends the terms to the appropriate translator. In this example, the terms are sent to an Arabic word translator (LanguageWeaver). The terms translated into Arabic are reverse translated into the original language, in this example English. The translated and reverse translated terms are returned to the GetTranslatedWordService for display to the learner. Figure 8 shows the translated and reverse translated terms. The terms on the far left are the terms selected from the “Expanded Term” page. The Arabic terms in the middle are the Arabic translations of the terms on the left. The terms on the right are the reverse translations of the Arabic terms.

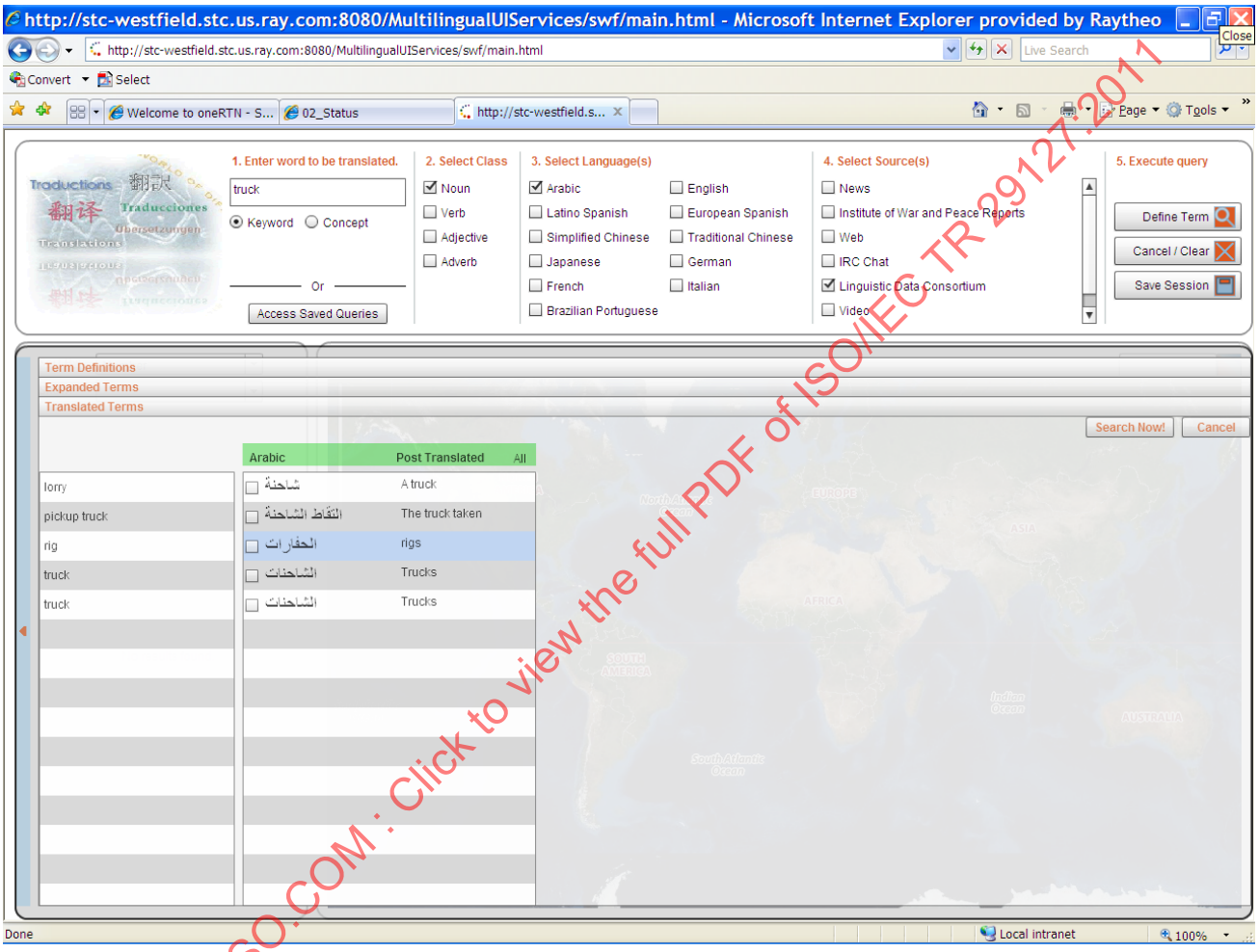

Figure 8 — Translated and Reverse Translated Terms

3.5 Selecting Appropriate Translations and Executing the Query

The learner selects the appropriate terms to use in the query. Please note there are errors in the translation of some of the terms. For example, “pickup” a type of truck was translated into Arabic as “taken”. This is an obvious miss-translation of what was semantically intended by the learner. The learner does not select this term for use in the query, as it would return information not useful to the learner. Also note the term “van” was translated into Arabic as “the”. This also is not the semantic intent of the learner for the term “van”. The learner does not select this term for use in the query, as it would return information not useful to the learner. Figure 9 shows the terms selected by the learner for the query.

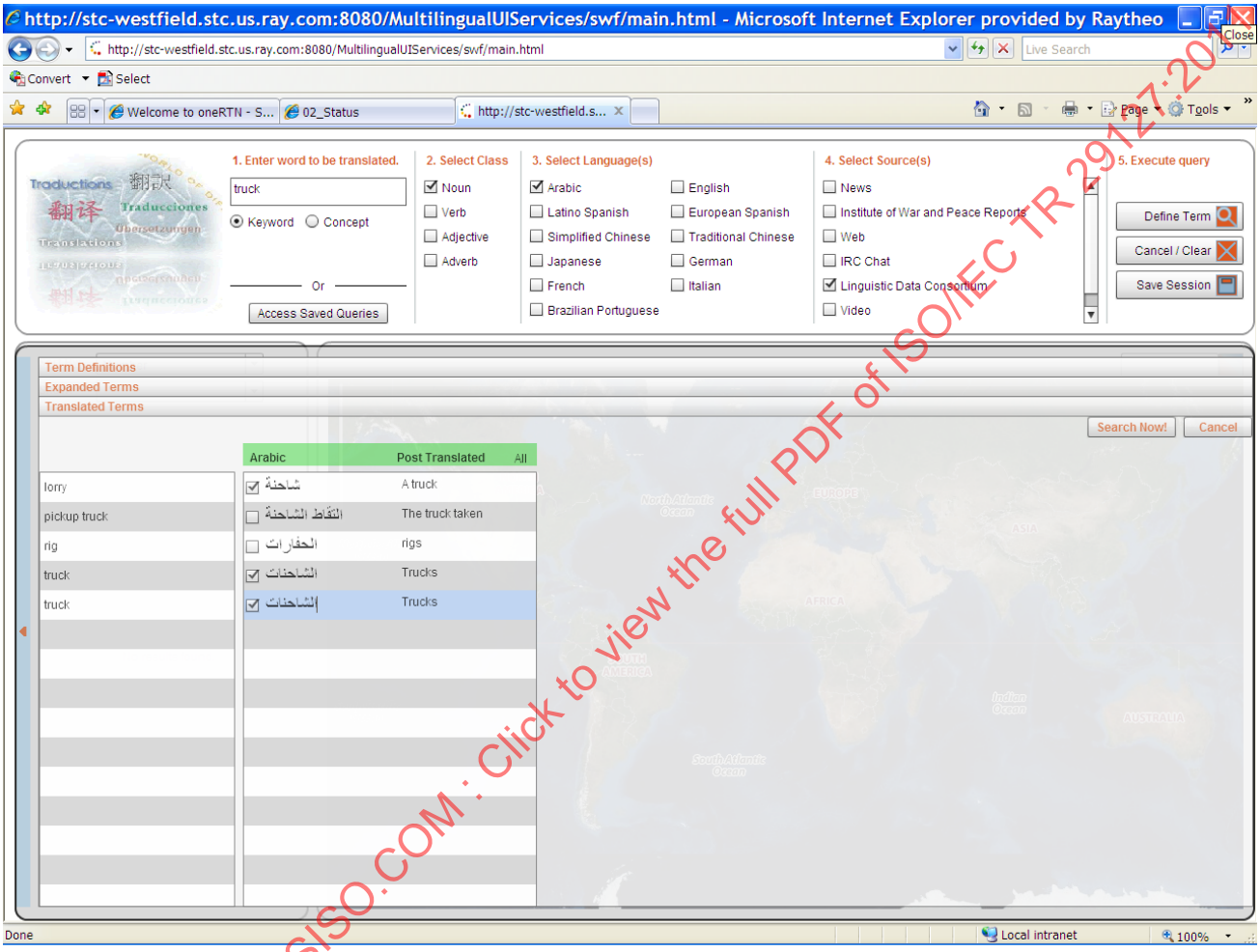

Figure 9 — Terms Selected for Query

The learner executes the query by selecting “Search Now” to the right of the translated and reverse translated terms. In this example, the SearchService places the Arabic terms into an SQL statement and sends the query to the source, in this example, the Linguistic Data Consortium (LDC) source. The LDC source consists of over 10,000 Arabic documents with a wide variety of topics.

3.6 Query Returns

The query returns are sent to extractors where summary information is obtained and then translated into the learner's language, in this example, English. The query return summary information is returned by the SearchService for display in the interface. Return summaries are displayed in the left side of the interface, ranked by return relevancy. Figure 10 shows the query returns ranked by relevancy.

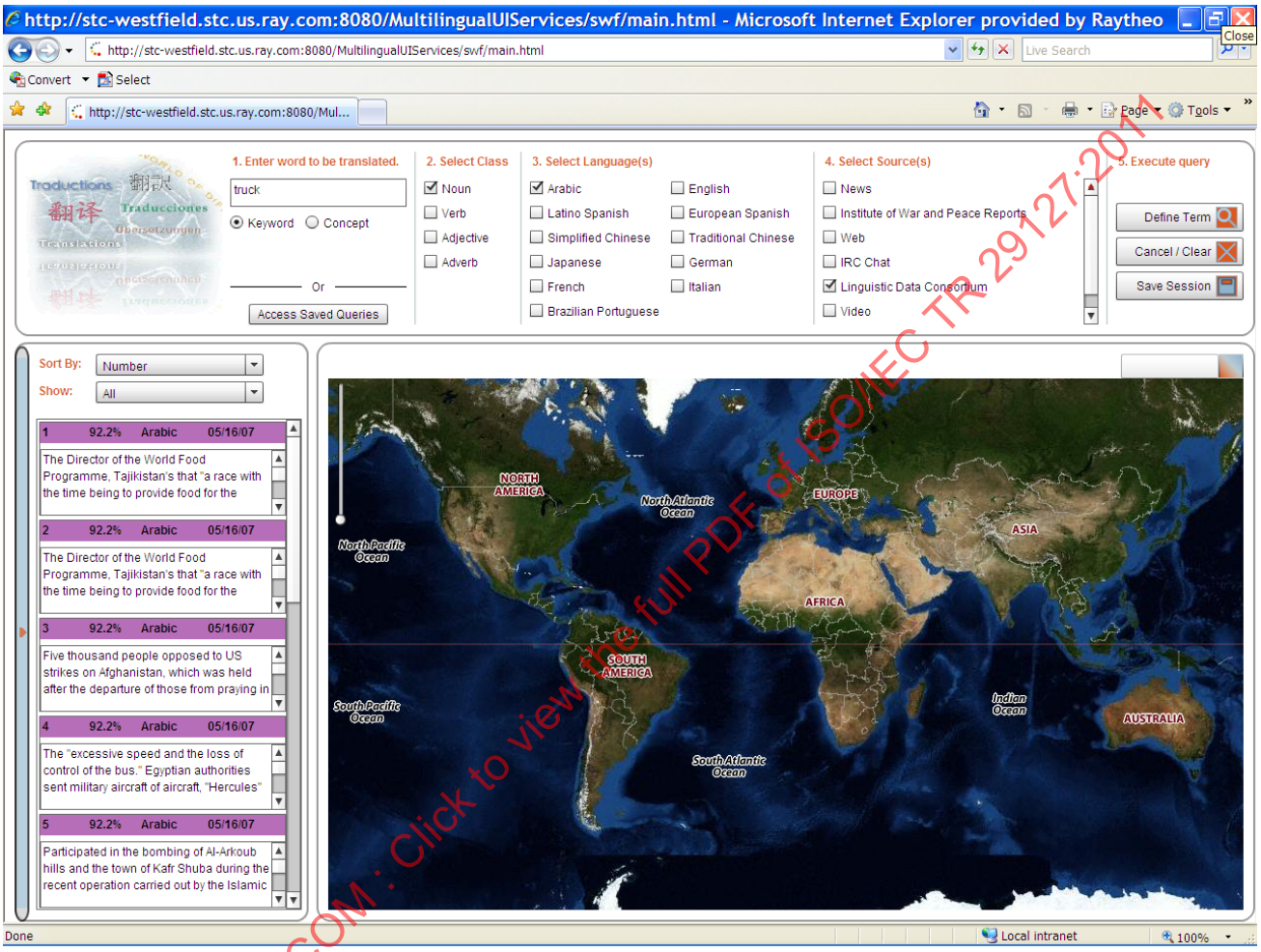

Figure 10 — Query Return Summaries

The learner selects a summary to retrieve the entire document. When the learner selects the appropriate summary, the GetSearchResultTextService, retrieves the entire Arabic document from the source, translates the document, and in this example, extracts additional information such as location data. The GetSearchResultTextService returns the Arabic document, and in this example, the English translation for display in the interface. The Arabic document the query was performed on is displayed on the left side of the document return section. The English translation is displayed on the right side of the display. Figure 11 shows the Arabic document and the English translation of the Arabic document.

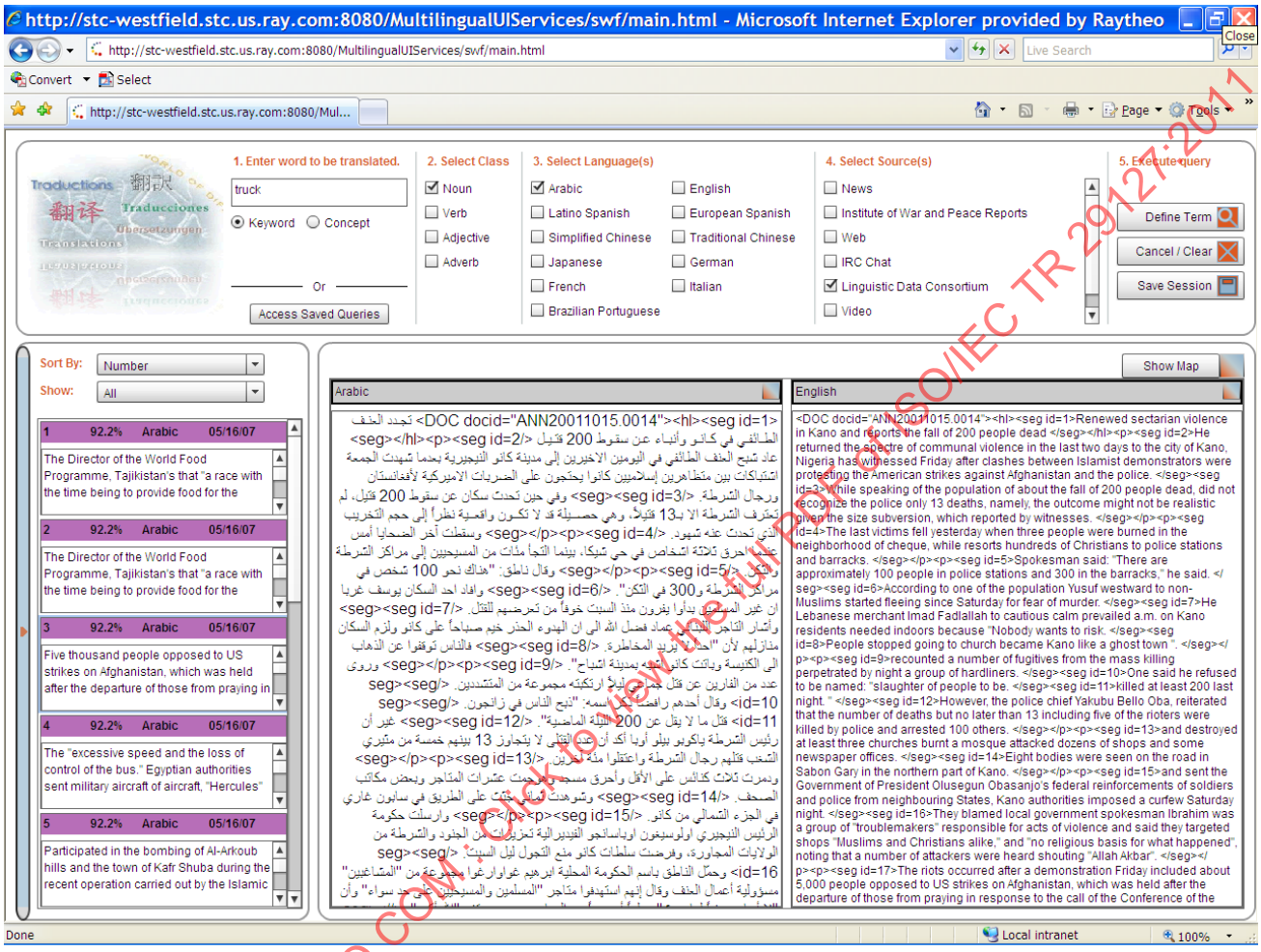

Figure 11 — Arabic Document and English Translation

In this example, character based locations are extracted from the Arabic document. The locations are converted to geolocation coordinates. In this example, the coordinates are sent to Yahoo maps for display purposes. The learner can view the map and plotted locations by selecting “Show Map” located above the English translation. Locations mentioned in the Arabic document are displayed on the map. The learner can mouse over the plot points to view the location name. Figure 12 shows the map with locations from the Arabic document plotted on the map. This facilities improved comprehension of the article’s place of origin or the location where the subject of the article can be found.

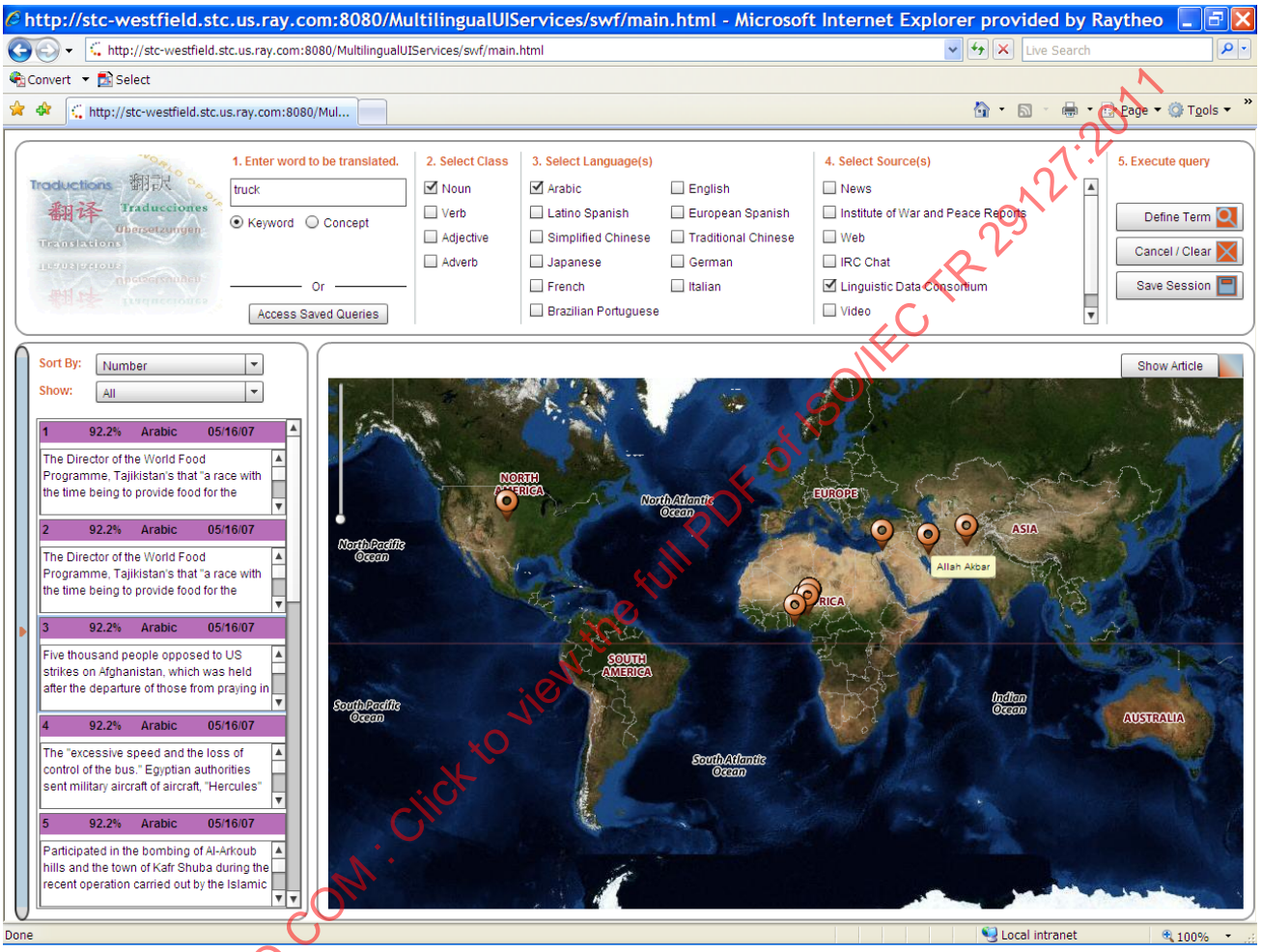

Figure 12 — Map With Plotted Locations

4 Components and Architecture of the SRQE Process

This clause contains the basic process diagram of the SRQE process and the J2EE architecture with the Java servlets and objects shown in a UML diagram format. The SRQE architecture is shown in Figure 13. Each of the architecture components are described in context to the SRQE process flow.

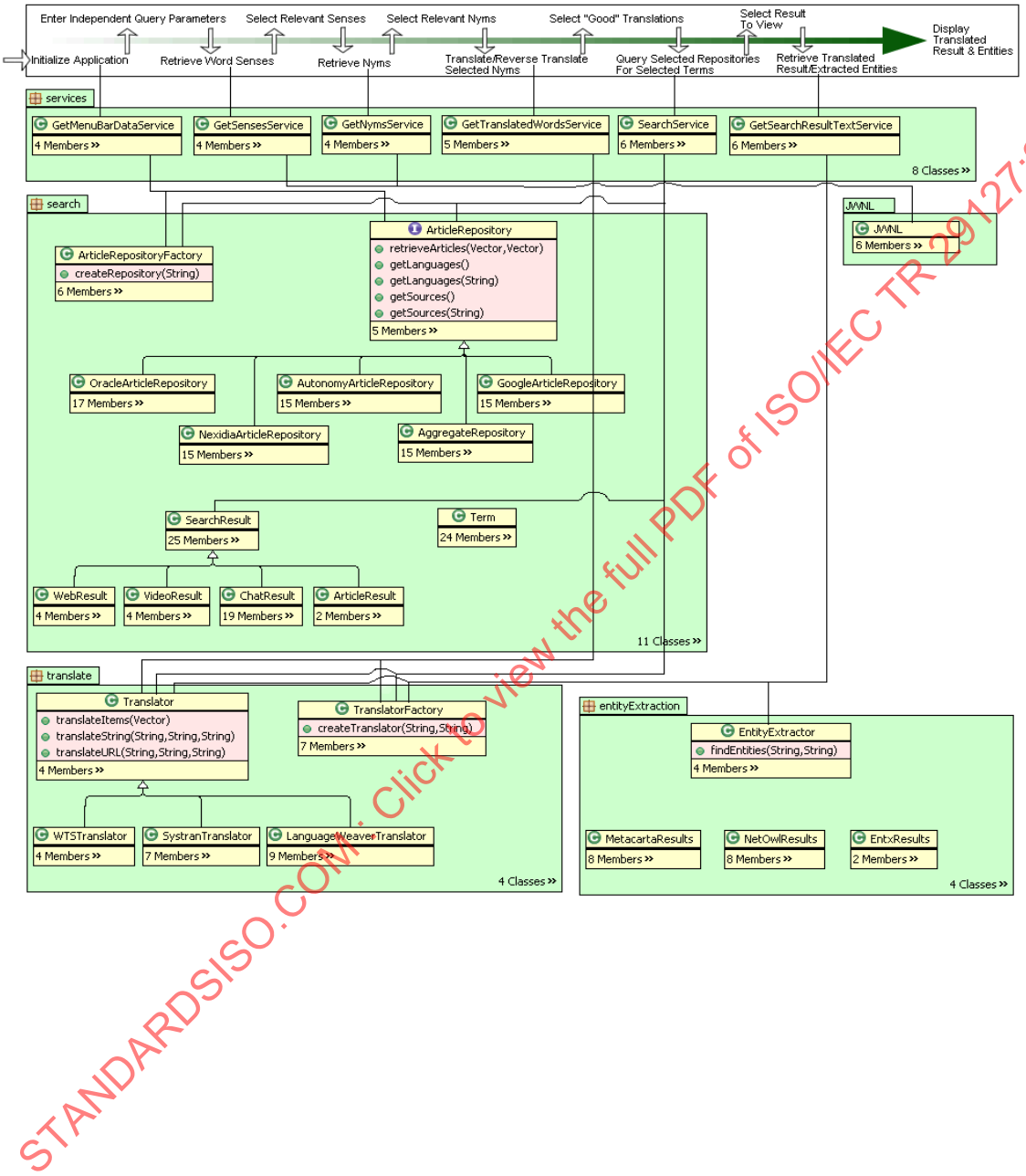

Figure 13 — SRQE Process Architecture

4.1 SRQE Process Flow

The SRQE process is a human-machine interactive process to perform cross-lingual queries. Figure 14 describes the SRQE process. Human inputs are shown in the white box above the green arrow. Java servlet functions are shown below the green arrow. A UML diagram showing the Java servlets is in a green box linked to functions in the white box.

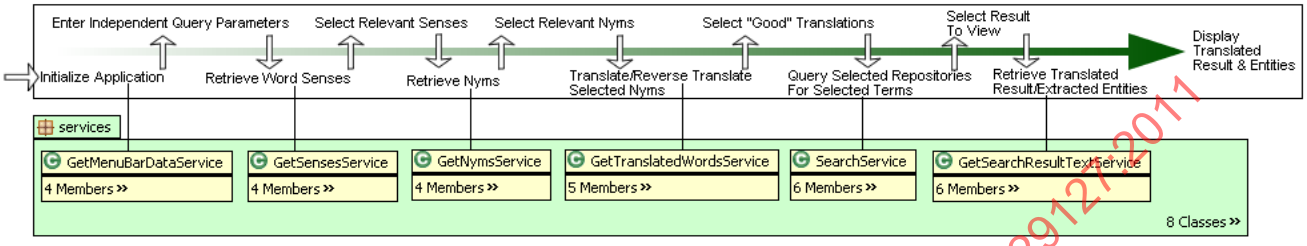

Bild 14 — SRQE Process

4.2 Repositories

A repository in the SRQE prototype is a generic concept for any service that contains indexed data that can be searched and retrieved by the SRQE prototype. A current example of a foreign language text document repository is an Oracle database utilizing Oracle Text.

Repositories are accessed using the ArticleRepository interface. ArticleRepository objects are created in a generic manner using the ArticleRepositoryFactory. Each repository provides access to any number of data sources (data sources refer to the original source of the data before the repository indexed it) and also provides the languages contained in those data sources. The ArticleRepository interface and the ArticleRepositoryFactory are shown in Figure 15.

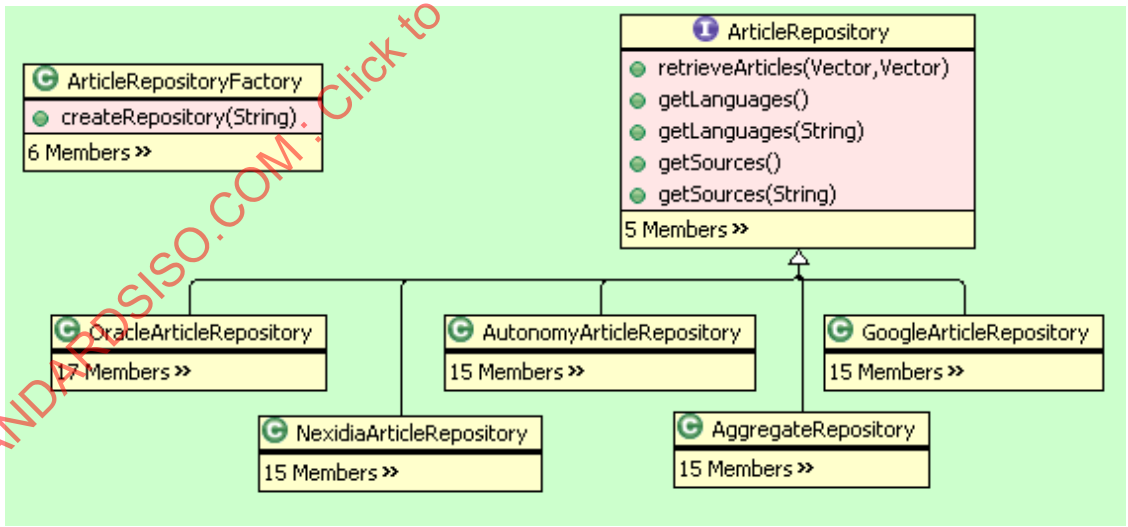

Figure 15 — ArticleRepository Interface and the ArticleRepository Factory

The ArticleRepository interface and the ArticleRepositoryFactory have two main functions. One function is to provide the UI with the sources and languages available for query for display to the learner. The GetMenuBarDataService uses the ArticleRepository interface and the ArticleRepositoryFactory. The repositories also provide the functionality to search the repositories based on learner approved terms and retrieve data contained in the repositories. The SearchService uses ArticleRepository interface and the ArticleRepositoryFactory.

Repositories need to have a specific ArticleRepository class written. Repositories can be implemented for use by modifying the search type mapping file used by the ArticleRepositoryFactory to construct the appropriate ArticleRepository object for the search type being performed.

4.3 Terms and Results

SRQE searches are based on the premise that given a list of expanded and translated terms, a list of search results can be obtained. The Term object encapsulates the original expanded term, the translation of that term, and the reverse translation of that term. The Term object is used by the GetSenses, GetNymsService, GetTranslatedWordService and the SearchService.

The search results can be of any SearchResult type, depending on the type of data that matches the search terms in the repositories available to be searched. The SearchResult object is used by the SearchService and the GetSearchResultTextService. SearchResult object and Term object are shown in Figure 16.

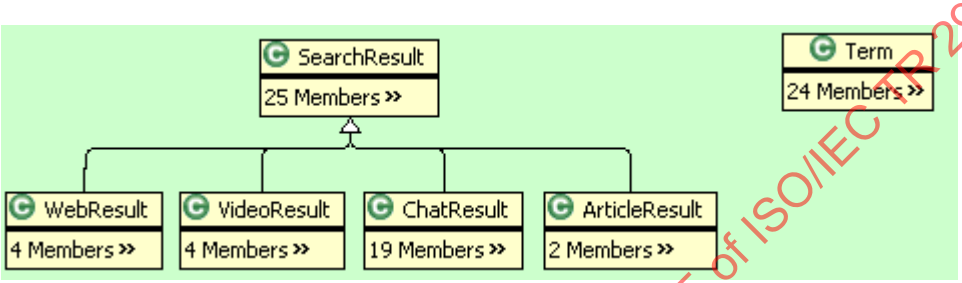

Figure 16 — SearchResult Object and Term Object

4.4 Translators

The concept of a translator is treated as a generic interface in the SRQE process architecture. Based on the original language and the target language, the appropriate translator object is constructed by the TranslatorFactory.

New translation tools can easily be added by writing a specific Translator class for the TranslatorFactory object. The language pair mapping file used by TranslatorFactory should be modified to use the new Translator class for all language pairs that are best handled by the new translation tool.

Both the TranslatorFactory object and the created Translator object created by the TranslatorFactory object, are used by the GetTranslationWordService, SearchService, and GetSearchResultTextService. The TranslatorFactory object and Translator object is shown in Figure 17.

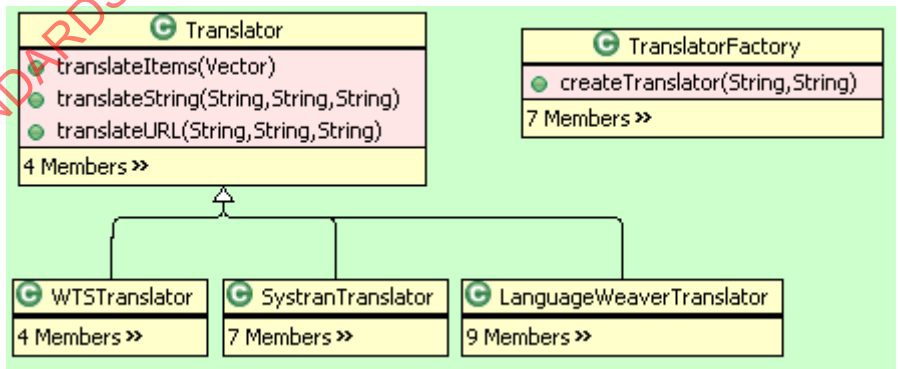

Figure 17 — TranslatorFactory Object and Translator Object

4.5 Entity Extraction

The SRQE process has the ability to incorporate any entity extraction tool that exposes an external API. The EntityExtractor class returns an extracted normalized Entity objects, regardless of the specific tool used to extract those entities. The entities are extracted from the native language text, if a tool is available to do so. Otherwise the translated text is used for extraction. The translated entities are then returned to the client and displayed to the learner. The GetSearchResultTextService uses the EntityExtractor object. The EntityExtractor along with sample extractors used in a prototype is shown in Figure 18.

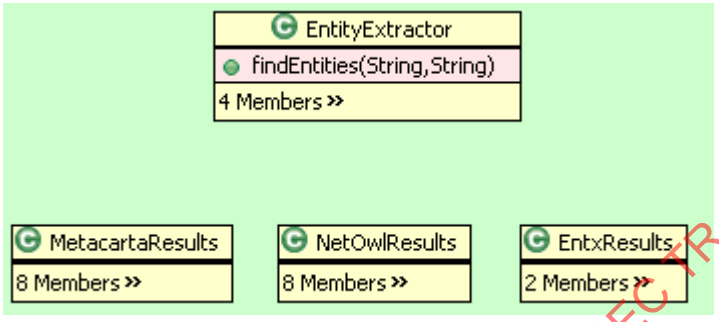

Figure 18 — EntityExtractor

4.6 Terminology for Query Searches

The SRQE process uses a terminology ontology, dictionary, knowledge base etc. to expand and select terms to include in a query. In most cross-lingual situations, there is often no word-to-word translation from one language to another. Additionally, there are often semantic ambiguities to be resolved. In both cases, term expansion helps. The SRQE process has the ability to incorporate any terminology repository that exposes an external API. For example, in Clause 3, an example implementation of the SRQE utilizing Princeton University WordNet is shown. In this example implementation, the Java WordNet Library (JWNL) is used by the GetSensesService and GetNymsService, using the JWNL API. The JWNL object is shown in Figure 19.

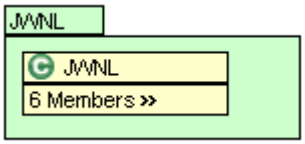

Figure 19 — JWNL Object

Annex A
(informative)

Potential Linkage to Current and Future ISO/IEC JTC1 SC 36 Technology Areas

The System Process and Architecture for Multilingual Semantic Reverse Query Expansion for LET identified in this Technical Report is used to query LET materials when the requestor does not understand or is not proficient in the language of the material available. It might be useful to identify other possible uses for the process and architecture identified in this Technical Report. The below matrix contains a list of potential linkages to current and future ISO/IEC JTC1 technology areas. This list is not meant to be exhaustive or definitive as information technology for LET purposes will evolve over time. The list is meant to be used as potential usage for the existing System Process and Architecture for Multilingual Semantic Reverse Query Expansion for LET as identified by this Technical Report. The Technical Report only identifies the existing process and architecture. The Technology Areas listed below are not required to utilize or incorporate the process or architecture contained in this Technical Report.

Table A.1 — ISO/IEC JTC1 SC36 Related Technology Areas

SRQE Processes/Services	SC36 Technology Areas	Linkage
Entire Process/All Services	Collaborative Technologies	Allow for querying of collaborative logs and or metadata of collaborative logs in multilingual contexts.
1.Entire Process/All Services 2. GetSenseService GetNymService GetTranslatedWordService	Collaborative Technologies ISO/IEC 19778-1 ITLET -- Collaborative technology-- Collaborative workplace -- Part 1: Collaborative Workplace Data Model ISO/IEC 19778-2 ITLET -- Collaborative technology-- Collaborative workplace -- Part 2: Collaborative Environment Data Model ISO/IEC 19778-3 ITLET -- Collaborative technology-- Collaborative workplace -- Part 3: Collaborative Group Data Model	Allow for semi-automated support of semantic disambiguation and/or translation of collaboration technology related data model elements and possibly their textual values, enabling semantic mappings of data model elements across languages and within a language.

SRQE Processes/Services	SC36 Technology Areas	Linkage
Entire Process/All Services	Collaborative Technologies	Allow for semi-automated support of semantic disambiguation and/or translation of collaboration technology related data model elements and possibly their textual values expressed in text enabling semantic mappings of data model elements across languages and within a language.
GetSenseService GetNymService GetTranslatedWordService Service	Metadata – In context of mapping semantically related terms that name metadata elements or that function as values assigned to these metadata in LET implimentations.	1. Allow for semi-automated support of semantic disambiguation and/or translation of metadata data elements and possibly their textual values enabling semantic mappings of metadata elements across languages and within a language within LET implementations. 2. Allow for semi-automated support of semantic disambiguation and/or translation of data associated with vocabularies and other interrelated sets of values.
Entire or part process/All or some services	Other ITLET activities e.g., language accessibility and human interface equivalencies, competencies, etc..	Allow for semi-automated support of semantic disambiguation of languages in the provision of multilingual equivalencies.

Annex B
(informative)

Patent Declaration Form for SRQE Process

Patent Statement and Licensing Declaration Form for ITU-T/ITU-R Recommendation | ISO/IEC Deliverable

Patent Statement and Licensing Declaration
for ITU-T/ITU-R Recommendation | ISO/IEC Deliverable
This declaration does not represent an actual grant of a license

Please return to the relevant organization(s) as instructed below per document type:

Director	Director	Secretary-General	General Secretary
Telecommunication Standardization Bureau International Telecommunication Union Place des Nations CH-1211 Geneva 20, Switzerland Fax: +41 22 730 5853 Email: tsbdir@itu.int	Radiocommunication Bureau International Telecommunication Union Place des Nations CH-1211 Geneva 20, Switzerland Fax: +41 22 730 5785 Email: brmail@itu.int	International Organization for Standardization 1 Chemin de la Voie-Creuse CH-1211 Geneva 20 Switzerland Fax: +41 22 733 3430 Email: patent.statements@iso.org	International Electrotechnical Commission 3 rue de Varembé CH-1211 Geneva 20 Switzerland Fax: +41 22 919 0300 Email: inmail@iec.ch

Patent Holder:

Legal Name: Raytheon Company

Contact for license application:

Name & Department: Philip P. Berestecki - Intellectual Property & Licensing Department

Address: 870 Winter Street
Waltham, MA 02451-1449

Tel.: 781.522.3039

Fax: 781.522.6465

E-mail: philip_p_berestecki@raytheon.com

URL (optional):

Document type:

☐ **ITU-T Rec. (*)** ☐ **ITU-R Rec. (*)** ☐ **ISO Deliverable (*)** ☐ **IEC Deliverable (*)**
(please return the form to the relevant Organization)

☐ **Common text or twin text (ITU-T Rec. | ISO/IEC Deliverable (*))** (for common text or twin text, please return the form to each of the three Organizations: ITU-T, ISO, IEC)

☒ **ISO/IEC Deliverable (*)** (for ISO/IEC Deliverables, please return the form to both ISO and IEC)

(*)Number: ISO/IEC 29127 (TR - Type 3)

(*)Title: System Process and Architecture for Multilingual Semantic Reverse Query Expansion for Learning, Education, and Training (LET)

Licensing declaration:

The Patent Holder believes that it holds granted and/or pending applications for patents, the use of which would be required to implement the above document and hereby declares, in accordance with the Common Patent Policy for ITU-T/ITU-R/ISO/IEC, that (check one box only):

[X] 1. The Patent Holder is prepared to grant a free of charge license to an unrestricted number of applicants on a worldwide, non-discriminatory basis and under other reasonable terms and conditions to make, use, and sell implementations of the above document.

Negotiations are left to the parties concerned and are performed outside the ITU-T, ITU-R, ISO or IEC.

Also mark here X *if the Patent Holder's willingness to license is conditioned on reciprocity for the above document.*

Also mark here X *if the Patent Holder reserves the right to license on reasonable terms and conditions (but not free of charge) to applicants who are only willing to license their patent claims, whose use would be required to implement the above document, on reasonable terms and conditions (but not free of charge).*

[ ] 2. The Patent Holder is prepared to grant a license to an unrestricted number of applicants on a worldwide, non-discriminatory basis and on reasonable terms and conditions to make, use and sell implementations of the above document.

Negotiations are left to the parties concerned and are performed outside the ITU-T, ITU-R, ISO, or IEC.

Also mark here __ if the Patent Holder's willingness to license is conditioned on reciprocity for the above document.

[ ] 3. The Patent Holder is unwilling to grant licenses in accordance with provisions of either 1 or 2 above.

In this case, the following information must be provided to ITU, and is strongly desired by ISO and IEC, as part of this declaration:

- granted patent number or patent application number (if pending);
- an indication of which portions of the above document are affected;
- a description of the patent claims covering the above document.

Free of charge: The words "free of charge" do not mean that the Patent Holder is waiving all of its rights with respect to the essential patent. Rather, "free of charge" refers to the issue of monetary compensation; *i.e.*, that the Patent Holder will not seek any monetary compensation as part of the licensing arrangement (whether such compensation is called a royalty, a one-time licensing fee, etc.). However, while the Patent Holder in this situation is committing to not charging any monetary amount, the Patent Holder is still entitled to require that the implementer of the above document sign a license agreement that contains other reasonable terms and conditions such as those relating to governing law, field of use, reciprocity, warranties, etc.

Reciprocity: As used herein, the word "reciprocity" means that the Patent Holder shall only be required to license any prospective licensee if such prospective licensee will commit to license its essential patent(s) or essential patent claim(s) for implementation of the same above document free of charge or under reasonable terms and conditions.

Signature:

Patent Holder	**Raytheon Company**
Name of authorized person	F. Kinsey Haffner
Title of authorized person	Vice President - Intellectual Property & Licensing
Signature	
Place, Date	Waltham, Massachusetts -- March 20, 2009

Annex C
(informative)

Summary on the Issue of Language Equivalencies

C.1 Introductory Notes to the ISO/IEC TR 29127 Annex C

The purpose of these "introductory notes" to Annex C is to update an original Canadian expert contribution on linguistic equivalencies, prepared in early 2007; and then to position it within the context of this TR on a "*System Process and Architecture for Multilingual Semantic Reverse Query Expansion (SRQE) for LET*".

The original contribution (SC36/WG7 N080 - 2007-02-20) was prepared for the members of WG7 and the project editors of what was then Part 8 of ISO/IEC 24751 {see 36N1666 for the text of the CD} on individual accessibility as one of several tools for the further development of what will now become Part 1 (Framework and Reference Model) of a new multipart ISO/IEC 20016 NP *"ITLET – Language Accessibility and Human Interface Equivalencies (HIEs) in e-Learning applications: Principles, Rules, and Attributes of Semantic Data"*. {See 36N1830} The primary set of requirements governing the development of this new multipart ISO/IEC 20016 standard, and especially its Part 1, is the *"UN Convention on the Rights of Persons with Disabilities*[2]*". (2006)*[3]

It is important that this contribution (and this Annex) be read with its original context in mind as this original document has only been marginally updated for this Annex C and any discrepancies and lack of smoothness in presentation are due to this difference in contexts and purpose between the two documents.[4]

---

[2] http://www.un.org/disabilities

[3] A key aspect of this *UN Convention* is its focus on the requirements that individuals have the right to be fully informed in order to be able to make decisions and make commitments. This requires that the semantics of the information presented at the human interface equivalent (HIE) level must be in a language that the individual is compatible in communicating.

[4] The scope, justification and purpose of ISO/IEC 20016 as found in 36N1830 (2009-06-01) are as follows:

> Scope:
>
> *This standard states the principles, rules, and attributes for semantic data, for specifying language accessibility and human interface equivalents (HIEs) in e-learning applications. It is structured to be able to support the requirements of applicable jurisdictional domains. Many jurisdictional domains have enacted legislation, regulations or policies that require equal access to education and information. In addition, there is the new UN Convention on Rights of Persons with Disabilities (2006). This standard can be used in meeting these requirements.*
>
> Justification and Purpose:
>
> *The language utilized in learning, education and training (LET) is determined by three key factors; namely:*

1) the language of the learner (apart from the learner wanting to learn another language);

In addition, the focus of SC36/WG7 N080 was to identify and analyze two approaches to language equivalences; translation theory and indexing theory (and thesauri construction) in particular. The reason for investigating these two approaches was that not only are they the two most prevalent means of establishing linguistic equivalences (within a single language, or among multiple languages), but the two approaches together analyze the multiple contexts involved in creating these equivalences. In understanding all of the concepts used in these two approaches, one is more aware the breadth, depth and width of the subject area concerned; which in turn helps us to provide better tools to achieve meaningful human linguistic equivalency. The approaches discussed below not only do they deal with language at the word (or lexical level), and then at the orthographical, phonological, morphological, and syntactic levels, but also take into account the sentence, phrase, document, inter-textual and extra-linguistic contexts, (e.g., including cultural impact). The study of these two approaches illustrates too the various inter-relationships of all aspects of language, i.e., the whole language system, as well as the relationships between other language systems. One aspect, however, that is not addressed in that neither of these two approaches, nor this TR, deal with the issue of multiple source languages and the establishment of equivalent meaning. This is where the multipart ISO/IEC 20016 comes in. It does this by introducing the concept of "semantic equivalence".

This Annex C aims to provide added information on the complexity of determining "linguistic equivalences". The aim of the SQRE is to provide a means to facilitate multilingual querying of recorded information using an approach which entails semantic equivalency matching. Understanding some of the complexity in the field allows one to appreciate more fully what the SRQE has accomplished. The SRQE as described and presented in this TR is one of many processes (and architectures) which allow for the development of linguistic equivalences, and it is a significant one in that it illustrates the ability to deal with multiple languages, different types of contexts, and various sets of recorded information. Clauses C.2 through C.4 below provide an edited version of document SC36/WG7 N080. Clause C.4 provides most of the "meat" for this discussion.

This Annex C has the following Clauses:

- C.1 Introductory Notes to the ISO/IEC TR 29127 Annex C
- C.2 Introduction and context of SC36/WG7 N080
- C.3 Purpose
- C.4 Working notes on "language equivalency"

---

2) *the needs of the learner with disabilities and anyone in a disabling context (this includes providing the semantics of the contents in the form of a human interface equivalents (HIEs) and doing so in a systematic and IT-facilitated manner); [Note: We already have both different forms of written representations of a language as well as in the form of symbols, glyphs, oral, pictorial, etc.]*
3) *the fact that the language of instruction (LOI), and thus the development of products is often governed by: (a) general rules governing use of an official language (or de facto language) of the jurisdictional domain in which the LET takes place; or, (b) a particular law or regulation of a jurisdictional domain which pertains to use of a language for LET purposes, i.e., as a "legally recognized language (LRN)".*

The key objective of this standard is to support language accessibility and facilitate multilingualism and do so from a perspective which incorporates both linguistic and nonlinguistic equivalents. Included here are non-written forms of communication such as oral, pictorial, etc.

Having a common IT-facilitated approach will: (1) not only (a) benefit individual users worldwide (doing so in respect and support of cultural diversity); (b) ensure that requirements of jurisdictional domains (at whatever level) can be supported in a very cost-effective and efficient manner; but, (2) also benefit suppliers of LET focused products.

Multilingual communications (whatever the IT platform utilized including the Internet) is already supported by existing technologies. Many ISO/IEC and ISO standards already exist (or are under development) whose contents can and will be utilized building blocks and integrated for this new LET standard.

The text below includes the complete text of the original SC36/WG7 N080 document; however, the following alterations were made:

- a new Clause C.1 “Introductory Notes to the Annex C” was added placing the original work in its new context;
- minor editorial and typographical changes were made;
- all references to ISO/IEC 24751-8 were changed to ISO/IEC 20016
- Table 2 from ISO 2788:1986 which illustrates monolingual thesauri lexical equivalencies was added;
- Clause 4 “List of terms and sources” was deleted;
- Clause 5 “Definitions by source”[5] was also deleted;
- several footnotes, where applicable, were added for clarification especially where the original context needed further explanation for the new Annex C context; and,
- all the references were updated and web-based links were added, wherever applicable.

C.2 Introduction and Context of SC36/WG7 N080

The purpose and scope of the new multipart ISO/IEC 20016 is that of:

1) supporting language accessibility as a key aspect in supporting both (a) the legal requirements of jurisdictional domains in this area; and, (b) the user, i.e., individual, needs, client-centred approach;

2) doing so through the concept/construct of human interface equivalents (HIEs);

with the focus being on communicating the semantics (meaning) of the recorded information being interchanged among the individuals and facilitating the use of information technology systems in support of the same.

As such, ISO/IEC 20016 focuses on supporting the availability, and thus the accessibility, of HIEs in whatever language a human being wishes to use or is capable of using in communicating semantics through recorded information. Thus, ISO/IEC 20016 is directed at providing concepts and methodology as tools in support of “semantic equivalency”.

This is quite different from “language equivalency” which is a field where extensive work has been done and for which many international (and national standards) exist. Key examples here are the “thesauri” standards[6].

Yet, there are many concepts, constructs, rules, definitions, etc., in the fields of translation and indexing which are applicable to language accessibility and HIEs, i.e., as “building blocks” or a “bridges”.

C.3 Purpose

The purpose of this expert contribution is to bring together in a single working (and reference) document for SC36/WG7 and its Project Co-editors, concepts related to “language equivalency” as found in international standards[7], and contributions to JTC1/SC committees of particular relevancy to

[5] The previous Clauses 4 and 5 provided detailed working notes, definitions and information required for the work on ISO/IEC 20016 which are superfluous to this current context and work on ISO/IEC 29127 TR.

[6] The three “thesauri” standards analyzed here are ISO 2788:1986, ISO 5964:1985 and ANSI/NISO Z39-19 (2005).

[7] The international ISO standards used as sources, i.e., ISO 2788 and ISO 5964, are available in both English and French versions. For the purposes of this Annex, only the English versions have been used. This has no bearing on the translation equivalence-based theories.